Hindawi
Complexity
Volume 2021, Article ID 5533884, 10 pages
https://doi.org/10.1155/2021/5533884

Research Article

Machine Learning-Based Multitarget Tracking of Motion in Sports Video

Xueliang Zhang[1] **and Fu-Qiang Yang**[2]

[1]*Department of PE and Art Education, Zhejiang Yuexiu University, Shaoxing, Zhejiang 312000, China*
[2]*School of Data and Computer Science, Shandong Women's University, Jinan, Shandong 250002, China*

Correspondence should be addressed to Xueliang Zhang; 20131102@zyufl.edu.cn

Received 30 January 2021; Revised 10 March 2021; Accepted 15 March 2021; Published 22 March 2021

Academic Editor: Wei Wang

In this paper, we track the motion of multiple targets in sports videos by a machine learning algorithm and study its tracking technique in depth. In terms of moving target detection, the traditional detection algorithms are analysed theoretically as well as implemented algorithmically, based on which a fusion algorithm of four interframe difference method and background averaging method is proposed for the shortcomings of interframe difference method and background difference method. The fusion algorithm uses the learning rate to update the background in real time and combines morphological processing to correct the foreground, which can effectively cope with the slow change of the background. According to the requirements of real time, accuracy, and occupying less video memory space in intelligent video surveillance systems, this paper improves the streamlined version of the algorithm. The experimental results show that the improved multitarget tracking algorithm effectively improves the Kalman filter-based algorithm to meet the real-time and accuracy requirements in intelligent video surveillance scenarios.

1. Introduction

Target detection and tracking is an emerging technology in computer vision, which has become the focus of many scholars in this field and is mainly used in intelligent video surveillance, traffic control, military surveillance, and other areas [1]. The implementation of target detection and tracking relies on a variety of technologies, such as digital image processing, artificial intelligence, etc. The content is to provide the database and prerequisites for subsequent pattern recognition, behaviour understanding, and image analysis through the detection and tracking function.

Network construction has developed rapidly in recent decades, and technology in various computer fields is developing rapidly, so people being eager to replace humans with computers to complete the boring panoply of work, information acquisition, and processing work is one of them [2]. Intelligent video surveillance can be completed to assist people to simplify the information acquisition process, extract key information, achieve information analysis and other functions, and intelligent video surveillance compared with traditional video surveillance, intelligent video surveillance can not only complete simple monitoring tasks but also can replace the human brain with a computer for information analysis and behaviour understanding [3]. On this basis, the intelligent surveillance video can also analyse and understand the movement behaviour of the target. Intelligent video surveillance is autonomous, preventive, and early warning and can accomplish tasks such as tracking specific people as well as vehicles, detecting the location of unexpected events, and daily monitoring, which has great practical value for modern society [4].

This series of intelligent video surveillance content can be divided into three major modules and motion target detection and identification, target tracking technology is the basic module of intelligent video surveillance, and the technology in the basic module is mostly derived from image processing [5]. The research in this direction improves the performance of moving target detection and tracking algorithms in video sequences in various aspects, such as robustness and real-time performance [6]. At this stage of intelligent video surveillance, facing real-time background

update and the effectiveness of moving target detection and tracking problems arising from mutual occlusion of multiple targets need to be solved and improved, so the research and improvement of moving target detection and tracking technology in video images are of great significance and will also have a positive economic impact on the future of intelligent video surveillance system industry [7].

Given the above background, this topic focuses on the study and exploration of motion target detection and tracking technology to provide the theoretical basis and technical support for intelligent video surveillance systems.

2. Related Research Work

Many foreign well-known universities and research institutions have conducted continuous and in-depth research on motion target tracking technology [8]. So far, great progress has been made in the field of motion target tracking technology. For example, Song et al. conducted an in-depth study of the interframe difference algorithm and proposed an effectively improved algorithm [9]. And a research group led by De et al. has been working on dynamic targets, and they improved the algorithm of the traditional background model [10]. Jiang et al. improved the background model algorithm and improved the background model building as an adaptive process [11]. The National Science Foundation has conducted in-depth research on detection and tracking techniques for targets in complex environments and has done a lot of work in this area [12]. Vehicle and motion tracking have been extensively researched and developed by a team of researchers at the University of Reading, UK [13]. Scholars at the University of Maryland have developed a real-time visual monitoring system that enables not only the localization of human body parts but also the tracking of multiple people [14]. Meanwhile, international academic journals and academic conferences have made important contributions to the development of motion target tracking technology.

Pourshamsi et al. proposed a region-based fully convolutional neural network to integrate the detection results with the tracker output to produce more tracker-friendly candidates for the problem of possible unreliable targets in the detector output [15]. Mittal et al. were the first to propose modelling the data association problem as a cost network flow with nonoverlapping constraints and solving the minimum-cost flow in this network to find the globally optimal trajectory association for multiple objectives [16]. The global optimal greedy algorithm, also modelled using a graph-theoretic approach, solves the data association using the K shortest path. By modelling the trajectory-detection association problem as a linear programming problem, Ghamisi et al. used the simplex method for solving and applying it to the tracking of multiple feature points in stereo vision [17]. Zhang et al. formulated the multiobjective tracking as a continuous energy function minimization problem and approximated the global optimal association solution by finding the strong local minima through the conjugate gradient method [18]. Yuan and Pu proposed to design energy functions to effectively mine the sparse epistasis model of detection results and POI algorithms combined with motion re-retrieval ideas in motion reidentification applications, GoogLeNet network was used to train on a large-scale motion reidentification dataset to obtain extraction parameters for motion epistemic features, and cosine distance was used to measure the similarity between different motion epistemic features in the testing phase, which effectively improved the performance of multitarget tracking [19].

In related research, people have done less research on the method adopted in this article, and the research in this article has certain advantages. This paper focuses on the visual multitarget tracking problem in video surveillance system, based on the study and analysis of related literature at home and abroad, the most representative, practical value and research significance of the motion target as the main tracking object in this paper, based on the demand of adaptive target number change and real time, this paper mainly researches the online multitarget tracking algorithm based on motion target. The purpose of this paper is to realize the fast identification and localization of multiple motion targets in a complex environment, to track multiple targets in the sequence at the same time, maintain the consistency of multiple target identities between frames accurately, and get the motion trajectory of multiple motion targets, and to lay the foundation for subsequent tasks such as abnormal warning, behaviour analysis, and traffic monitoring.

3. Research and Analysis of Machine Learning for Motion Multitarget Tracking in Sports Video

3.1. Machine Learning Algorithm for Multiobjective Tracking. Digital image processing consists of a series of information acquisition and processing techniques in images and videos such as denoising, enhancement, recovery, segmentation, feature extraction, and target recognition. In this paper, we study the target detection and tracking algorithms under surveillance video, and the preliminary research work involves image denoising, image enhancement, and morphological transformation in image preprocessing. These image preprocessing techniques are briefly introduced and analysed in the following.

Median filtering uses a nonlinear digital filtering method, and both it and the homomorphic filtering method can enhance the image. Median filtering can improve the problem of image blurring brought by linear filtering under some conditions [20]. It is implemented by first sampling the input signal and discriminating the result of the sampling. When the discriminated result is representative of the signal, the values in the observation window are sorted, and the middle value of the sorted series is taken as the output; then, a new sample is obtained, and finally, the above operation is cycled. Median filtering has a good suppression effect for pretzel noise and has a good removal effect for the high-frequency part in Fourier space. Since the grey value of the high-frequency component in the image signal changes

rapidly, median filtering can play a role in eliminating some of the components. It also has a relatively good effect in the suppression of edge blurring occasions that preserve the edge characteristics:

$$G(x, y) = Med\{f(x + k, y + l), (k, l) \in W\}. \quad (1)$$

Gaussian filtering is a linear smoothing filter that can effectively smooth images and suppress Gaussian noise. The template coefficients in Gaussian filtering are varied according to the distance from the centre of the template, and the mean value is used as the output value.

Zero-mean discrete Gaussian filter function is

$$G(x, y) = \exp\left(-\frac{x^2 + 1}{\sigma^2}\right). \quad (2)$$

Two-dimensional zero-mean defeated 6-s function filter idle number is

$$G(x, y) = \exp\left(-\frac{x^2 + y^2 + 1}{2\sigma^2}\right). \quad (3)$$

The general Gaussian template is $3 * 3$ or $5 * 5$, and the following is the distribution of the weight values, respectively:

$$\frac{1}{32} \times \begin{bmatrix} 2 & 4 & 2 \\ 4 & 8 & 4 \\ 2 & 4 & 2 \end{bmatrix} \quad \frac{1}{556} \times \begin{bmatrix} 2 & 8 & 14 & 8 & 2 \\ 8 & 32 & 52 & 32 & 8 \\ 14 & 52 & 82 & 52 & 14 \\ 8 & 32 & 52 & 32 & 8 \\ 2 & 8 & 14 & 8 & 2 \end{bmatrix}. \quad (4)$$

The specific implementation is based on the previous state sequence, the optimal estimation of the next state of the system through the state equation and the observation equation. Because there are often disturbances such as noise in the system in the observed data, the optimal estimation can also be regarded as a filtering process. The optimal value of the present moment is obtained by combining the measured value of the current moment with the measured value of the previous moment and the noise error to find the optimal solution and update it, and similarly, the values of the velocity and position of the next moment can be obtained by iteration. The equations underlying the Kalman filter are as follows:

$$\begin{aligned} x_i &= A_{i*(i-1)}x_i + w_i, \\ z_i &= H_{i*(i-1)}x_i + v_i. \end{aligned} \quad (5)$$

For a general neural network, each pixel of the image is usually connected to each neuron in the fully connected layer, while a convolutional neural network connects each hidden node to only a local region of the image, thus reducing the number of parameters to train. In the convolutional layer of a convolutional neural network, the weights corresponding to the neurons are the same, and the number of trained parameters can be reduced because the weights are the same. Suppose that in a target detection system, A, B, C are used, and good results are achieved, but at this time you do not know which one of A, B, C plays a role, so you keep it A, B, remove C, and experiment to see the role of C in the entire system. Shared weights and biases are also referred to as convolutional kernels or filters. Since the images to be processed are often larger than large, the most important thing is to get effective image features without analysing the original image during the actual processing. Therefore, a simple convolutional neural network structure can be used like the idea of image compression for convolutional operations, where the image is resized by a downsampling process, as shown in Figure 1.

Theoretically, the fully connected feedforward neural network has rich feature representation capability, with the help of which it can be well used for the analysis of image images. However, there are several problems in practical use, such as the difficulty in adapting to the variability of images and the complexity of computation. Convolutional neural networks effectively solve the problems encountered by full connectivity in terms of local connectivity, weight sharing, and downsampling. The local connection is such that each neuron is not connected to every neuron in the previous layer. The use of local connectivity effectively reduces the number of parameters in the training process. Weight sharing allows a group of neuron connections to share weights, which also reduces the number of parameters and speeds up training. Downsampling uses pooling to reduce the number of samples per layer of the neural network and to improve the robustness of the model. For tasks related to image processing, convolutional neural networks increase the speed of training and ensure processing results by retaining feature parameters and reducing unnecessary parameters.

The convolutional layer reduces the connectivity of neurons by noncomplete connectivity, thus reducing the computational complexity, but the number of neurons is not significantly reduced, the dimensionality of subsequent computations is still high, and the overfitting problem can easily occur. To solve this problem, a pooling layer for pooling operations is introduced, also known as a downsampling layer. The downsampling layer can greatly reduce the dimensionality of the features, thus reducing the computational effort and avoiding the overfitting problem.

Although the tracking targets are different, they should all have some commonalities, which need to be learned by the network. However, training with tracking data is difficult because the same object, which is the target in one video sequence, may be the background in another video sequence. The targets in each video sequence are completely different, and various challenges arise, such as occlusion, deformation, and rotation. Many existing network models are mainly for tasks such as target detection, classification, and segmentation, and because they must divide many classes of targets, these networks are large and increase the computational burden. In the tracking problem, the network needs to be divided into only two categories: target and background. Usually, the tracked targets are small, so a very large network is not needed to solve the tracking problem.

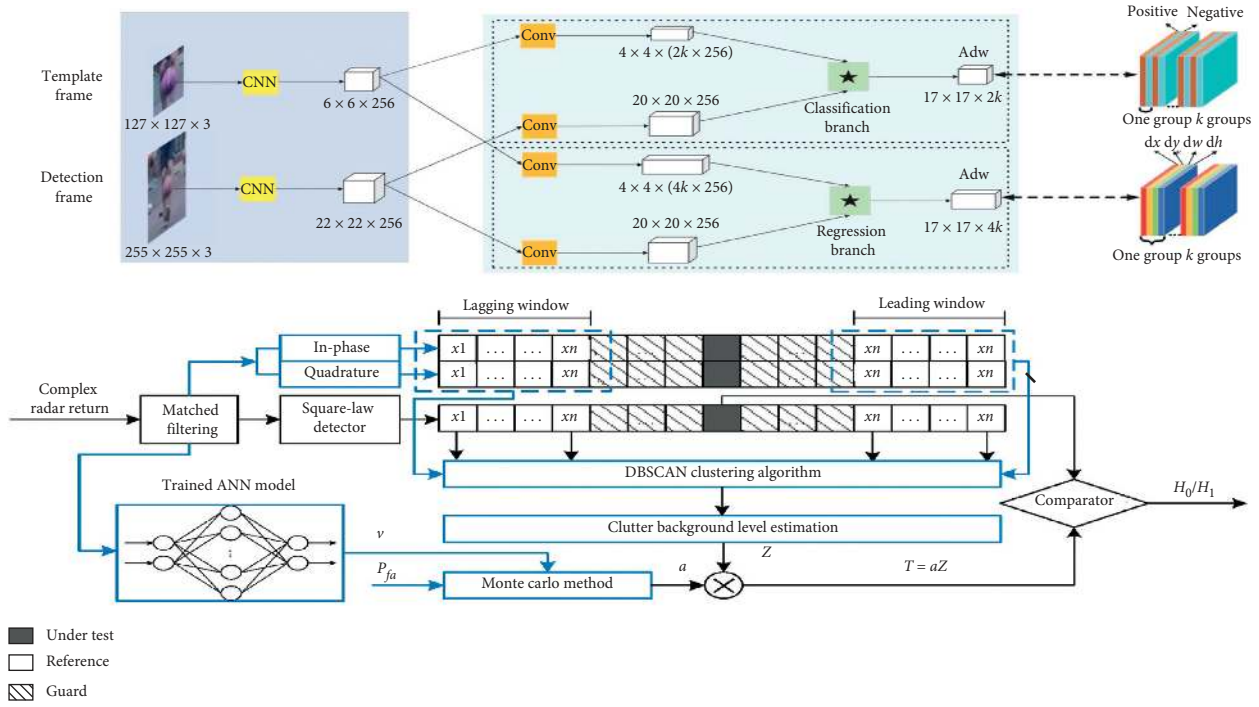

Figure 1: Network structure of machine learning algorithm for multitarget tracking.

MDNet is proposed to address the abovementioned points of the current situation by learning the commonality of these targets through a multidomain learning network structure.

In the detection of moving targets, there are different noises in the video image sequences, and the binary images obtained after thresholding usually have breaks and holes in the target area, which are processed using open and closed operations to improve the detection accuracy.

The essence of the open operation is to erode and expand the image, which eliminates burrs and scatter in the image. The closed operation is the opposite of the open operation in terms of implementation steps and can be used to fill in the hole effects that exist in the image.

Take A as the target object and B as the structure element ($3 * 3$ or $5 * 5$). The open operation on A with structural element B is denoted as

$$A \cdot B = (A \otimes B) \oplus B. \tag{6}$$

Most of the visible light in nature can be composed of three fundamental colours, and the RGB colour space is also constructed based on the principle of three fundamental colours, with the three components of the RGB colour space representing the x-, y-, z-axes, with the origin representing black and the dotted endpoints along the square in the figure representing white. Any colour light D can be composed of R, G, and B in different proportions:

$$D = R(p) + G(p) + B(p). \tag{7}$$

From the schematic diagram of the two-frame differential method, we can see that the algorithm has the characteristics of simple implementation steps, a small number of operations, and small requirements for system implementation. Because the core idea in the algorithm is to use the neighbouring two frames to do image differencing, so the camera is subject to fewer light and shadow interference factors; more suitable for the actual scene is the warehouse class of burglary video surveillance, but at the same time by the core idea of the algorithm, the algorithm is easy to contrast between the images to a lesser degree, and the difference is not obvious in the case of the resulting difficulty to extract the feature value of the moving target in the image. Since the time difference between two frames is too small and the sampling effect of the difference algorithm between two frames is related to the movement speed of the moving target, when the movement speed is small or even the moving target is stationary, the phenomenon of not detecting the moving target can occur:

$$D_k(x, y) = \left| f_{k+1}(x + k, y + l) - f_k(x, y) \right|. \tag{8}$$

The core idea of the average background method is to calculate the mean and standard deviation of each pixel as the background model:

$$\text{Average}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \left| f_{i+1}(x + k, y + l) + f_i(x, y) \right|. \tag{9}$$

When tracking motion targets with complex shapes, the target model created by kernel tracking cannot accurately represent the target, and the silhouette tracking method can describe the adaptive shape. The silhouette tracking and kernel tracking algorithms are the same in idea; both use the initial frame to build the target model. Silhouette tracking extracts the target contour based on the established target model and then uses contour matching to search the target area of the subsequent video frames.

Common silhouette tracking methods are divided into two categories: contour tracking methods and shape matching methods. The silhouette tracking method applies to the case of partial overlap of target silhouettes between consecutive frames, and the energy function and state-space model are used to estimate the position information of the target silhouette in the current frame. The contour tracking algorithm can adjust the centre of gravity and size of the contour according to the changing appearance of the moving target: the shape matching method is to search for the target shape by matching the established target model shape to the current video image. All three tracking algorithms have their applicable occasion conditions, and now the three methods are compared and analysed; see Table 1.

3.2. Multiobjective Scheme Design for Motion in Sports Video. A good multiobjective tracking dataset often determines a good or bad tracking algorithm. To have a fair and objective comparison with other algorithms, it is required that the dataset be complete and rich, contain all kinds of special cases, and can reflect the accuracy and robustness of the algorithm. A complete multitarget tracking dataset should contain (1) different viewpoints, for example, SORT-based target tracking has a high tracking accuracy for fixed viewpoints under low-altitude surveillance video, but a low accuracy when the video is changed to a moving viewpoint for in-vehicle video; (2) target scenes with different densities; if there are too few targets, then it cannot reflect the algorithm's handling of target occlusion, target deformation, and similar target interference; (3) different light intensity; the same target in computer vision, the pixel value of the corresponding position in a rainy day and a sunny day will produce a large difference; we can also use this difference to evaluate the robustness of an algorithm; (4) different video frame rate; the video frame size is sensitive to the relevant parameters of the corresponding model of the multitarget tracking algorithm, and similarly, we can also use this difference to evaluate the robustness of an algorithm. The cost of our calculation method is about 80% of other studies, and the results can be increased by 5%.

To make a comprehensive and fair comparison of different algorithms, this paper uses a publicly available dataset as a benchmark dataset for the performance measurement of multitarget tracking algorithms [21]. The dataset is the dataset used to measure the criteria of multitarget tracking methods in the multitarget tracking series. The dataset of this class contains a rich set of scenarios to meet the requirements of various situations mentioned above. The dataset contains a total of 14 video sequences, of which 7 video sequences form the training set and the remaining 7 sequences form the test set, each of which is very different from the other. It may be a top view, a flat view, or a top view. The photos are taken by high-resolution cameras, and each frame contains multiple detection frames on average, and each sequence contains multiple target numbers.

Generative models more widely used in image super-resolution, face recognition and generation, image complementation and style conversion, etc., while the application of generative models is still in its infancy in scenes with less detailed image features and more background interference like multitarget tracking. For this kind of image processing problems with complex interference information, a low percentage of positive sample pixels in the field of view, and inconspicuous target features, it is usually difficult for the generative model to fully learn the distribution of key data in the input samples, and the complex scene information has a large impact on the accuracy of the generative model sample authenticity determination [22]. The multitarget tracker may encounter the problem of drifting, misfollowing, or losing the identity due to the change of identity caused by the deformation of the target in the process of tracking specific targets [23–25]. In the online tracking process, the stability and accuracy of the tracker can be effectively improved if more diverse samples (e.g., samples of the same target in multiple views and multiple motion states) can be generated based on a limited number of samples of a specific target in the history frame. Therefore, this paper proposes a generative model based on conditional variational self-encoder-conditional generation adversarial network and analyses and adjusts the target to be tracked in a multitarget tracking environment to generate multiview samples for a specific target to expand the training space of that target to enhance the performance of the tracking algorithm.

Variational self-encoder (VAE), shown in Figure 2, is a type of self-encoder and is an improved structure for generative tasks. In essence, it adds Gaussian noise to the standard self-encoder structure on top of the hidden variable of the output of the encoder (the encoder used to calculate the mean of the distribution), so that the decoder can be more stable to noise changes during the training process and adds KL scatter as a regular term to constrain the zero mean for the output of the decoder; In addition, an additional encoder is added to dynamically adjust the strength of Gaussian noise by calculating the variance of the input distribution, i.e., during the training process of the decoder, when the reconstruction error is larger than the KL scatter, the noise is appropriately reduced to reduce the fitting difficulty, while when the reconstruction error is smaller than the KL scatter, the noise is appropriately increased to increase the fitting difficulty so as to improve the decoder's ability to generate samples.

The principle of variational self-encoder is based on variational inference, and the overall structure is a complete probabilistic inferential model, which can estimate the edge probability of the input image and maximize the edge probability to adjust the parameters of the model for training; it can also estimate the Gaussian posterior probability of the hidden variable to realize the decoding process from the hidden variable to the generated sample. Based on the theory of variational inference, the lower bound estimate can be obtained by calculating the parameters of the variational lower bound, and this estimate can be optimized using the standard stochastic gradient descent method; i.e., backpropagation can train the model. This also means that the two-part structure of the variational self-encoder can be

Table 1: Comparison of the analysis of three tracking methods.

Tracking method	Advantage	Disadvantage
Point tracking	Suitable for translational movements with small targets	When the target shape is obscured, point tracking
Nuclear tracking	When the target contour is a geometric rigid body, the effect is better	Inability to effectively track a small number of feature points' illumination and target morphology changes have high sensitivity
Silhouette tracking	It has a significant effect on tracking nonrigid bodies with complex shapes	High requirements for target model establishment

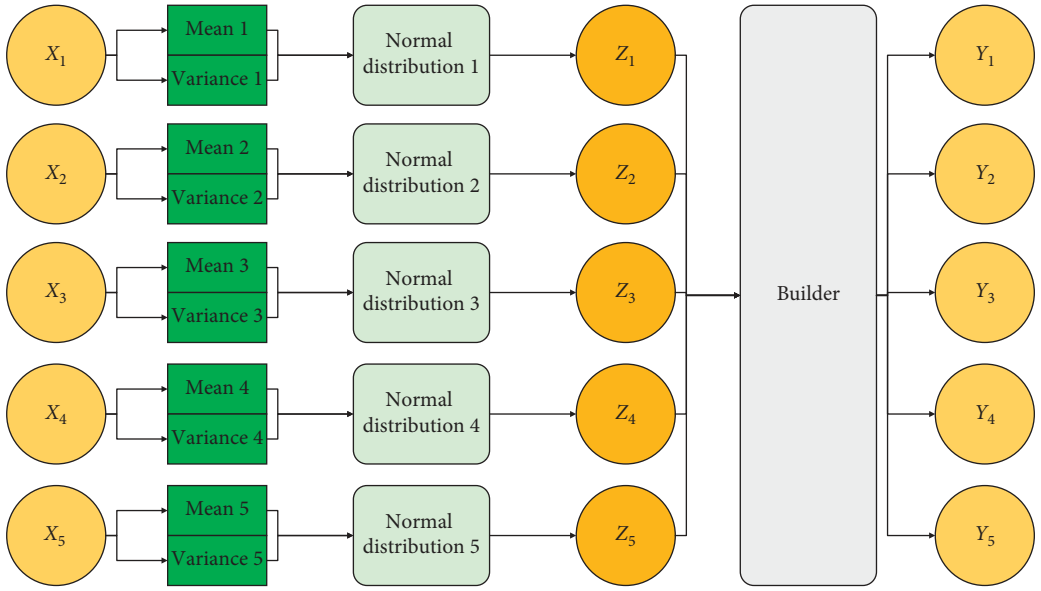

Figure 2: Working principle of the variable division self-encoder.

reconstructed using a deep learning model to perform most of the tasks in computer vision.

4. Analysis of Results

4.1. Validation of Multitarget Tracking Data Enhancement Effect. The ultimate goal of the proposed method in this paper is to be able to compensate for the lack of data diversity in multitarget tracking tasks; therefore, the proposed model is used to validate two approaches to multitarget tracking, respectively, a single-target tracking enhancement-based tracking method and a twin neural network-based tracking method. The fundamental difference between the two approaches is the update mechanism.

The twin neural network-based tracking method does not have an online update process during tracking, avoiding the cumulative error that occurs during constant updating. However, this method relies on offline training, and the training process is directed at the edges containing the target to be tracked not the global information of a single frame. For multitarget tracking tasks, the diversity of datasets available for training is not strong. When the trained twin neural networks are used for tracking, the frequent occurrence of interactions and occlusions due to the uncertain number of targets in the field of view can greatly affect the similarity metric accuracy of the single-sample trained twin neural networks. The model proposed in this paper expands the dataset during the training process and uses the generative model to generate multiangle and multipose states for samples of the same identity, which makes the twin neural network more robust to complex situations, as shown in Figure 3.

The above two methods are optimized using their corresponding data generation methods, and the final performance is verified on the MOT Challenge 17 dataset, and the results are shown in Figure 3. From the experimental results, it can be seen that after expanding the data diversity with the generation model proposed in this paper, it can effectively reduce the impact of both methods by possible deformation and occlusion, as shown in the reduction of False Negative (FN) by 3.1% and 4.8%, respectively, and the reduction of Identity Switch (ID Sw.) by 22.5% and 18.3%, respectively. Due to the enrichment of training space, the recognition rate and long-time tracking effect of the two methods are also improved, as shown by the improvement of the index by 0.4 and 0.7, respectively. In summary, the proposed method can enrich the diversity of multitarget tracking data as a data generation method, which can effectively improve the overall performance of multitarget tracking.

The multitarget tracking algorithm based on spatial information association uses motion estimation and prediction, combines spatial distance metric and tracking gate restriction

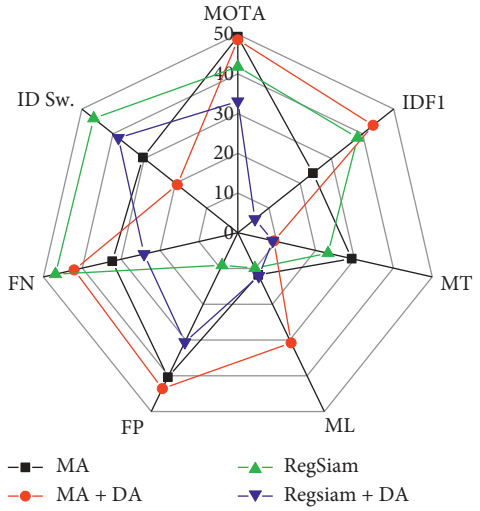

Figure 3: Effect of multitarget tracking data enhancement on tracking performance.

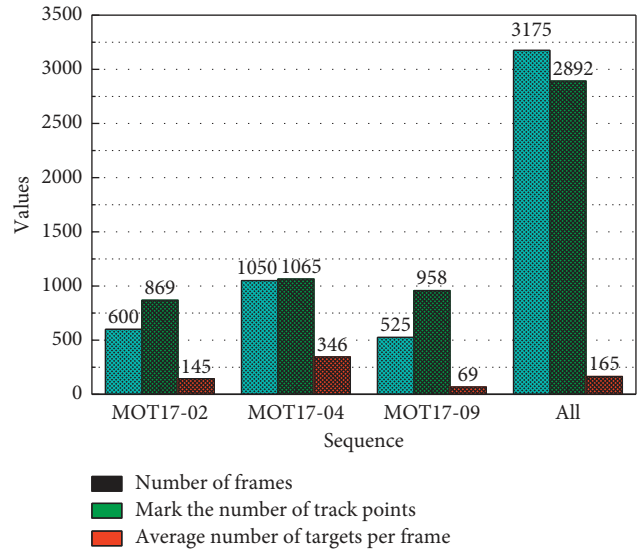

Figure 4: Multitarget tracking experimental test set parameters.

to construct the cost matrix of data association, and uses the Hungarian algorithm to solve the tracking matching result under the minimum association cost; the experimental environment is shown in Figure 4. Three surveillance videos under the fixed camera in MOT17 are used as the test set, and the motion target is with visibility greater than 30%. The specific parameters of the three videos are shown in Figure 4.

To quantitatively compare the effects of different detection algorithms on the multitarget tracking effect, this paper compares the accuracy, precision, and several tracking quality evaluation indexes for the above three groups of tracking results, where the arrow next to the name of each index indicates the expected trend of the tracking algorithm on the index, and the downward arrow indicates that the smaller the index is, the better the index is, and the upward arrow indicates that the larger the index is, the better the index is. In the experiments of this chapter, the confidence threshold of both the Faster RCNN and the improved YOLO detection results are uniformly set to 0.8, and the threshold of DPM is set to −1. The evaluation metrics of the three groups of tracking results are obtained as shown in Figure 5.

From Figure 5, it can be seen that the multitarget tracking results based on the improved YOLO algorithm in this paper perform better in four indexes, MOTA (accuracy), MT (most tracking), ML (most lost), and missed followers (FN), indicating that the detection and tracking algorithm in this paper not only reduces the number of missed followers but also improves the stability of target tracking. The tracking results based on the Faster RCNN detection algorithm are more accurate than those of DPM and YOLO, mainly because the Faster RCNN algorithm based on the two-step method has higher accuracy in the regression of the target bounding box, which leads to a higher intersection ratio between the successfully tracked target and the real trajectory bounding box.

This comparison experiment is designed to test the impact of association cost metrics on data association algorithms by using Euclidean distance between target frame

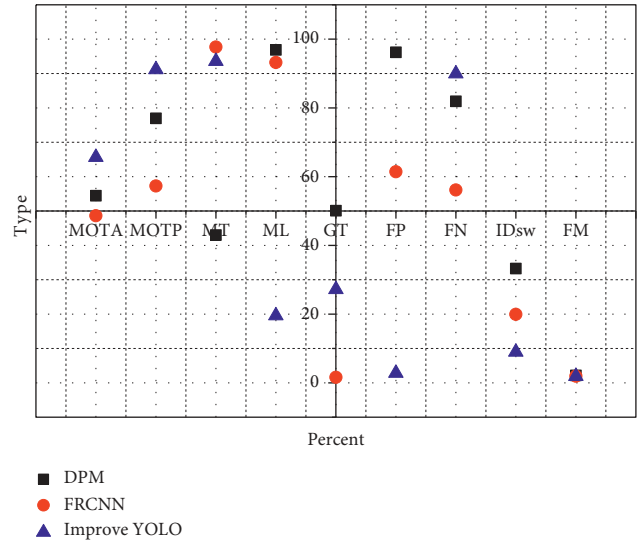

Figure 5: Comparison of motion multiobjective tracking result indicators with different detection algorithms.

centroid locations, intertarget frame intersection and merge ratio, and the proposed tracking-gate-based intersection and merge ratio of target frames, respectively. In this paper, the distance metric for multitarget data association uses the cross-merge ratio between the detection result located in the tracking gate and the predicted state bounding box, which more than doubles the tracking accuracy compared to using the Euclidean distance between the target centre positions. The change to the intersection ratio as the distance metric is the most useful, and the metric not only has better matching performance compared to the Euclidean distance metric but also applies to targets of different sizes with convenient parameter settings and high portability. The addition of the tracking window qualification based on the motion state estimation makes the number of target identity transformations in the tracking process decreases and the number of

false alarms decreases. However, due to the addition of the tracking gate qualification, the number of misses and tracking fragments in the tracking results also increases slightly.

4.2. Performance Test Results Analysis of Motion Multitarget Tracking System in Sports Video. In this section, the proposed method in this paper is compared with the methods in the literature. Firstly, experiments are conducted separately using the combination scheme of different viewpoints to observe the tracking performance improvement effect of multiple viewpoints. Then a fair comparison with existing methods is performed to analyse the advantages and shortcomings of the proposed method in this study, and the comparison results are shown in Figure 6.

The input–output evaluation of the experimental results is consistent with the evaluation system in the literature, using the same inputs as well as the same GTs, which can ensure the fairness of the comparison. As shown in Figure 6, the performance of the tracking system in this study achieved good results, although it did not exceed the literature; however, the 100% MT as well as the lower IDs has exceeded the other comparison methods. In the more difficult intensive motion scenarios, the best results were achieved by the method proposed in this paper. This indicates that Markov optimization models play a key role in dealing with dense scenes. It can also be found from Figure 6 that the method in this study makes better use of multiview data for tracking, i.e., the tracking performance using data from all three views is better than that using data from only two views in all three experimental sequences, which follows the practical physical meaning and is consistent with the original purpose of multiview research.

In Figure 7, the three subplots of each plot represent the simultaneous tracking result plots for the master view and the two slave views, respectively. It can be seen from the plots that tracking using the information from multiple viewpoints enables a more accurate estimation of the target trajectory. It can also be found that the same targets are correctly assigned the same tracking trajectory markers in different viewpoints.

A Markov random field data association optimization model based on cross-view coupled trajectory fragments is introduced, which has a new potential function enhancement method to effectively associate the fragments of trajectory fragments caused by intensive motion. In this paper, the cross-view coupled trajectory fragments are obtained by a data fusion method based on image mutual information. This method can calculate the spatial position relationship between cross-view 2D trajectory fragment pairs considering both position and motion information and corrects the position coordinates of stumped and deviated target detection data in dense motion scenes using the human key point detection method. Using PETS 2009 experimental dataset for modular and system-level experiments, respectively, the experimental results demonstrate the enhancement of image mutual information method and MRF optimization method in the process of

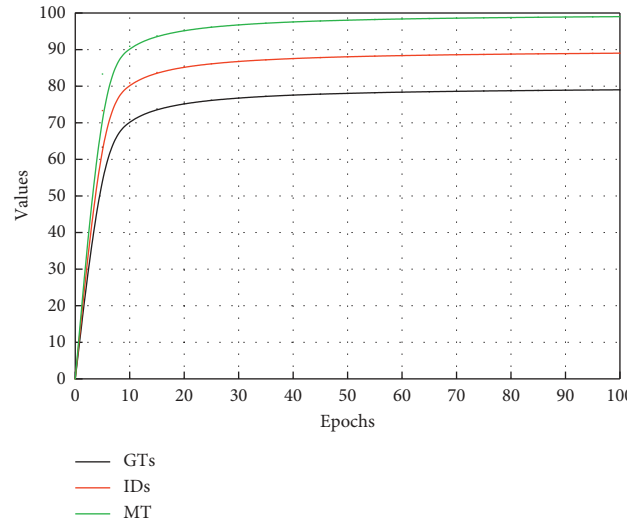

Figure 6: Tracking system performance comparison results.

data fusion and data association. Meanwhile, the tracking method in this study has good multiview multitarget tracking performance when compared with the current research methods.

To test the running speed of the multitarget tracking algorithm in this chapter, and the change of the overall running speed of the detection and tracking algorithm when the number of targets in the scene changes, three static surveillance videos in MOT17 were tested and the results were counted, and the average time consumed by the algorithm is shown in Figure 8.

From Figure 8, it can be shown that the average time consumed by the tracking algorithm in this chapter is only about 9 ms, which is fast. On the other hand, the average number of targets in the three test sequences is different, resulting in a larger dimension of the cost matrix to be solved in the scenario with a larger number of targets in the data association process, which in turn increases the computational effort of the multitarget tracking segment, but since the tracking segment itself accounts for a smaller proportion of the overall algorithm operation time, the difference in the average frame rate is limited, which in turn indicates that the number of targets in the scenario has a limited impact on the tracking algorithm and has limited fluctuations in the computation time. This paper optimizes the traditional classifier for cosine similarity to achieve the extraction and effective matching of target depth epistatic features. Then, the fusion and tracking strategy of apparent feature similarity and spatial similarity in multitarget tracking is proposed.

The construction of target similarity measure and association cost is the core of data association based on Hungarian arithmetic. The target bounding box intersection and ratio based on tracking gate used in this chapter is more than double in tracking accuracy compared to using Euclidean distance between target centre locations, and the change to intersection and ratio as a distance metric plays the biggest role, which is not only applicable to multiscale targets and easy to set threshold parameters but also has

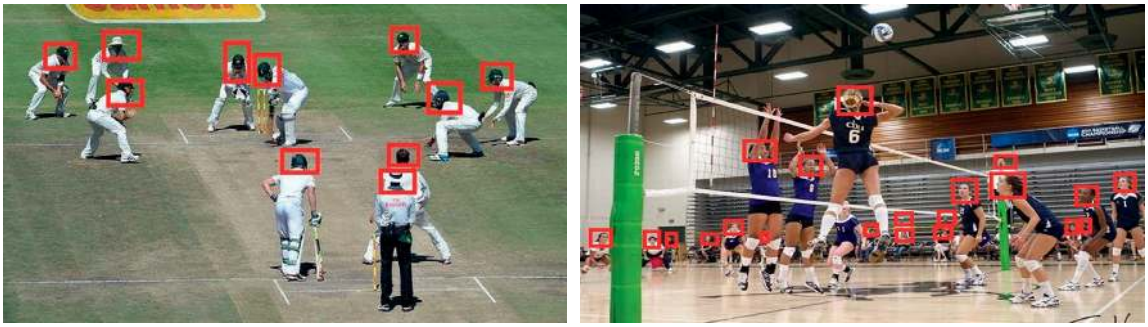

Figure 7: Trace result graph.

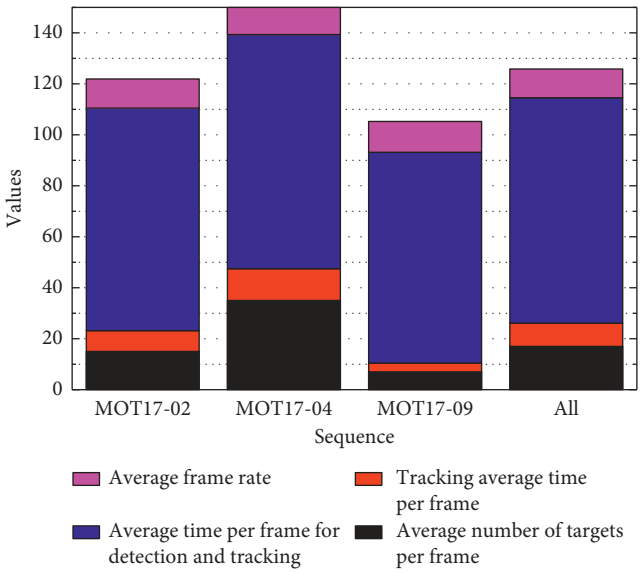

Figure 8: Test elapsed time and average frame rate of multitarget tracking algorithm on MOT17 dataset.

better tracking performance compared to Euclidean distance metric.

Compared with the latest related research, our research results show that our efficiency and accuracy have improved by nearly 5%. On this basis, the tracking gate limit based on the motion state estimation is introduced to further reduce the number of target identity transformations and the number of false alarms in the tracking process. However, the method also has shortcomings, and the number of missing targets and tracking fragments in the tracking results is slightly increased due to the increase of the limits.

5. Conclusion

This paper focuses on the motion multitarget tracking technology in video surveillance and deeply investigates the multitarget online tracking problem in video surveillance from four aspects: motion target detection, motion state estimation, data association, and epistemic feature modelling. The construction of a feature pyramid network for multiscale target prediction and the combination of multiscale training mechanism effectively improve the quality of multiscale motion target detection in complex surveillance scenes and combine with the clustering optimization of the a priori frame parameters to effectively reduce the rate of missed detection and false detection of motion targets. Based on the target detection results, the overall design of the detection-based multitarget tracking scheme in this paper is designed in combination with video surveillance application scenarios, and a multitarget tracking algorithm that fuses apparent feature metrics with spatial similarity is proposed. This paper proposes a multitarget tracking algorithm that fuses the apparent features of the target with spatial information. When the cross-entropy loss function based on the classifier is used for training, the obtained feature distribution shows radial characteristics and does not match the distance of the commonly used feature similarity metric, and for this problem, this paper optimizes the traditional classifier for cosine similarity to achieve the extraction and effective matching of target depth epistatic features. Then, the fusion and tracking strategy of apparent feature similarity and spatial similarity in multitarget tracking is proposed. To address the shortage of effective update mechanism of target features in the field of multitarget tracking, a sparse update mechanism with intertarget occlusion judgment is proposed to effectively prevent tracking drift in the case of intertarget occlusion. The effectiveness of this paper's method is verified by the experimental results on three sets of surveillance video sequences.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] J. Hülsmann, J. Traub, and V. Markl, "Demand-based sensor data gathering with multi-query optimization," *Proceedings of the VLDB Endowment*, vol. 13, no. 12, pp. 2801–2804, 2020.

[2] A. Belhadi, Y. Djenouri, J. C.-W. Lin, and A. Cano, "Trajectory outlier detection," *ACM Transactions on Management Information Systems*, vol. 11, no. 3, pp. 1–29, 2020.

[3] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: model-based, AI-

based, or both?" *IEEE Transactions on Communications*, vol. 67, no. 10, pp. 7331–7376, 2019.

[4] A. B. Hamida, A. Benoit, P. Lambert, and C. B. Amar, "3-D deep learning approach for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4420–4434, 2018.

[5] A. Farasat, G. Gross, R. Nagi, and A. G. Nikolaev, "Social network analysis with data fusion," *IEEE Transactions on Computational Social Systems*, vol. 3, no. 2, pp. 88–99, 2016.

[6] H. A. Pierson and M. S. Gashler, "Deep learning in robotics: a review of recent research," *Advanced Robotics*, vol. 31, no. 16, pp. 821–835, 2017.

[7] L. Li, K. Ota, and M. Dong, "Deep learning for smart industry: efficient manufacture inspection system with fog computing," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 10, pp. 4665–4673, 2018.

[8] A. Jalali and H. Farsi, "A new steganography algorithm based on video sparse representation," *Multimedia Tools and Applications*, vol. 79, no. 3-4, pp. 1821–1846, 2020.

[9] H. Song, J. J. Thiagarajan, P. Sattigeri, and A. Spanias, "Optimizing kernel machines using deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 11, pp. 5528–5540, 2018.

[10] S. De, L. Bruzzone, A. Bhattacharya, F. Bovolo, and S. Chaudhuri, "A novel technique based on deep learning and a synthetic target database for classification of urban areas in PolSAR data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 1, pp. 154–170, 2018.

[11] L. Jiang, L. Yan, Y. Xia, Q. Guo, M. Fu, and K. Lu, "Asynchronous multirate multisensor data fusion over unreliable measurements with correlated noise," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 53, no. 5, pp. 2427–2437, 2017.

[12] H. Wu, Z. Zhang, C. Jiao, C. Li, and T. Q. S. Quek, "Learn to sense: a meta-learning-based sensing and fusion framework for wireless sensor networks," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8215–8227, 2019.

[13] Z. Zhao, X. Wang, and T. Wang, "A novel measurement data classification algorithm based on SVM for tracking closely spaced targets," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 4, pp. 1089–1100, 2019.

[14] M. A. Al-Jarrah, A. Al-Dweik, M. Kalil, and S. S. Ikki, "Decision fusion in distributed cooperative wireless sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 1, pp. 797–811, 2019.

[15] M. Pourshamsi, M. Garcia, M. Lavalle, and H. Balzter, "A machine-learning approach to PolInSAR and LiDAR data fusion for improved tropical forest canopy height estimation using NASA AfriSAR Campaign data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 10, pp. 3453–3463, 2018.

[16] N. Mittal, U. Singh, R. Salgotra, and B. S. Sohi, "An energy efficient stable clustering approach using fuzzy extended grey wolf optimization algorithm for WSNs," *Wireless Networks*, vol. 25, no. 8, pp. 5151–5172, 2019.

[17] P. Ghamisi, R. Gloaguen, P. M. Atkinson et al., "Multisource and multitemporal data fusion in remote sensing: a comprehensive review of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 1, pp. 6–39, 2019.

[18] H. Zhang, X. Zhou, Z. Wang, H. Yan, and J. Sun, "Adaptive consensus-based distributed target tracking with dynamic cluster in sensor networks," *IEEE Transactions on Cybernetics*, vol. 49, no. 5, pp. 1580–1591, 2019.

[19] X. Yuan and Y. Pu, "Parallel lensless compressive imaging via deep convolutional neural networks," *Optics Express*, vol. 26, no. 2, pp. 1962–1977, 2018.

[20] D. Nada, M. Bousbia-Salah, and M. Bettayeb, "Multi-sensor data fusion for wheelchair position estimation with unscented Kalman Filter," *International Journal of Automation and Computing*, vol. 15, no. 2, pp. 207–217, 2018.

[21] V. Radu, C. Tong, S. Bhattacharya et al., "Multimodal deep learning for activity and context recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, pp. 1–27, 2018.

[22] Q. Zhou and Y. Zheng, "Long link wireless sensor routing optimization based on improved adaptive ant colony algorithm," *International Journal of Wireless Information Networks*, vol. 27, no. 103, pp. 241–252, 2019.

[23] S. Xia, D. Peng, D. Meng et al., "A fast adaptive k-means with no bounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1, 2020.

[24] B. Yong, W. Wei, K. C. Li et al., "Ensemble machine learning approaches for webshell detection in Internet of things environments," *Transactions on Emerging Telecommunications Technologies*, 2020.

[25] J. S. Almeida, P. P. Rebouças Filho, T. Carneiro et al., "Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques," *Pattern Recognition Letters*, vol. 125, pp. 55–62, 2019.